CGIAR Research Program on
Climate Change, Agriculture and Food Security (CCAFS)

Guidance for handling different types of Data
Video Transcript

October 2013

Data Management Guidelines by Statistical Services Centre, University of Reading is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Permissions beyond the scope of this license may be available at www.reading.ac.uk/ssc.

These materials were produced for and with funding from the Climate Change Agriculture and Food Security Research Program of the Consultative Group on International Agricultural Research (CGIAR).

Introduction

This video focuses on different types of data collected that are usually stored as variables, and uses parts from a questionnaire to illustrate, although concepts covered in this video are relevant for any type of data collection activity.

It is important to start thinking about the details introduced in this video during the design of the activity tools; this should ensure that the data collection, entry and analysis will run more smoothly.

Identifier Variables

The identifier data usual consists of several variables that are commonly concatenated and referred to as the unique identifier or primary key.

Identifier variables can be numeric, text or alphanumeric depending on the context.

The identifier variables and primary key must uniquely identify each measurement unit in the activity (whether this is an individual or plot).

All datasets must have a variable or set of variables that uniquely identifies each record.

The identifier variables relating to individuals or households must be anonymised.

Numeric or Continuous Data

Numeric or continuous data is the result of questions or measurements that naturally give a number response, such as number of household members, yield measurements, plot size, age, amount of fertiliser applied etc.

Different types of numeric data will require slightly different formats to account for different levels of precision. For example number of individuals or age should always be recorded as a whole number with no decimal places or fractions; whereas when measuring a response such as plot size it is usually appropriate to include a number of decimal places

When numeric data are the result of measurement the units should be specified in advance, or be recorded in a separate text column if they are not consistent across all observations (plot sizes in hectares, acres and local units).

Additional numerical fields can be derived as part of the transition from raw to primary data.

Categorical Data

Categorical data, or factors, are questionnaire or measurement responses that fall into categories, for example respondent gender: male/female

Where categorical data is being stored numerically, each of the categories should be assigned a code and an accompanying data dictionary must always be provided to decode the information, for example 1 = Male, 2 = Female.

Code lists (from within a data dictionary) can be recycled across questions if the same categories are used in more than one question, for example using a single list of occupations the primary and secondary occupations of an individual can be recorded.

Free Text Data

Free text data is text data that cannot be coded into categories, such as respondent name, or data resulting from open ended questions such as 'give your opinion…', 'do you have any further comments…', or enumerator comments.

Free text should be entered exactly as seen in the original source; no text should be paraphrased, however punctuation and accents will need to be removed. If free text may need to be used for any analysis, for example if the frequency of respondents mentioning a key phase or word is of interest, it is useful for the data to have been entered either all in upper case or all in lower case.

Multiple Response Data

Multiple response data is usually the result of responses being a list, for example name the main types of crop grown on your land, what livestock do you keep etc.

The first option for storing this data, is for the number of variables (columns) being the maximum number of responses given by any individual. The first response given by a respondent is entered into the first variable, the second response into the second variable and so on.

The second option is to create a variable for each unique response given, and indicate whether each crop was specified by the respondent using Yes/No or coded 1/2 responses. (This is usually preferred by the analyst; however it can result in a large number of variables).

Missing Data

Avoid collecting missing data – for categorical variables ensure the group responses cover all options: for example when asking about highest level of education obtained include a category for 'no education' so these responses are not collected as missing. You can also add 'NA' and 'Other' codes with space to add the details.

When collecting numeric data ensure the value for missing cannot be a plausible response, for example use -99, or a number 10 times that which could be considered a response such as 999. Try to use consistent missing codes, this will assist the enumerator and reduce the number of mistakes during data collection.

The document that corresponds to this video includes additional advice on the types of data introduced in this video as well as discussing date data, general variable and dataset specifications, and formats for storing images, and videos.

Appendix I – CCAFS Data Management Support Pack

This document is part of the CCAFS Data Management Support Pack produced by the Statistical Services Centre, University of Reading, UK. The following materials are available in the pack:

0. Data Management Strategy
    a. CCAFS Data Management Strategy
1. Research Protocols
    a. Writing Research Protocols – a statistical perspective
    b. Preparation of Research Protocols – Good Practice Case Study
    c. What is a Research Protocol, and how to use one (Video & Transcript)
    d. Details of what a Research Protocol should contain (Video & Transcript)
2. Data Management Policies & Plans
    a. Creating a Data Management Plan
    b. Data Management Plan (Video & Transcript)
    c. Example Data Management Activity Plan
    d. Example Consent Form
3. Budgeting & Planning
    a. Budgeting & Planning for Data Management
    b. ToR Data Support Staff
    c. Budgeting & Planning (Video & Transcript)
4. Data Ownership
    a. Data Ownership and Authorship
    b. Template – Data Ownership Agreement
    c. CCAFS Data Ownership & Sharing Agreement
    d. Data Ownership & Authorship (Video & Transcript)
5. Data & Document Storage
    a. Creating and Using a DDS
    b. DDS Introduction – (Video & Transcript)
    c. DDS Organisation – (Video & Transcript)
    d. DDS Ownership – (Video & Transcript)
    e. Introduction to Dropbox – (Video & Transcript)
6. Archiving & Sharing
    a. Archiving & Sharing Data
    b. Data and Documents to Submit for Archiving – a checklist
    c. MetaData
    d. Archiving & Sharing (Video & Transcript)
    e. Metadata (Video & Transcript)
    f. CCAFS HBS Questionnaire
    g. CCAFS HHS Code Book
    h. CCAFS Training Manual for Field Supervisors

7. CCAFS Data Portals
    a. Portals for CCAFS Outputs
    b. AgTrials Summary
    c. CCAFS-Climate Summary
    d. DSpace Introduction
    e. Introduction to Dataverse (Video & Transcript)
    f. Creating a Dataverse (Video & Transcript)
    g. Dataverse Study Catalogue
    h. CCAFS Dataverse (Video & Transcript)
8. Data Quality & Organisation
    a. Data Quality Assurance
    b. Guidance for handling different types of Data
    c. Transition from Raw to Primary Data
    d. Data Quality Assurance (Video & Transcript)
    e. Guidance for handling different types of data (Video & Transcript)
    f. Transition from Raw to Primary Data (Video & Transcript)